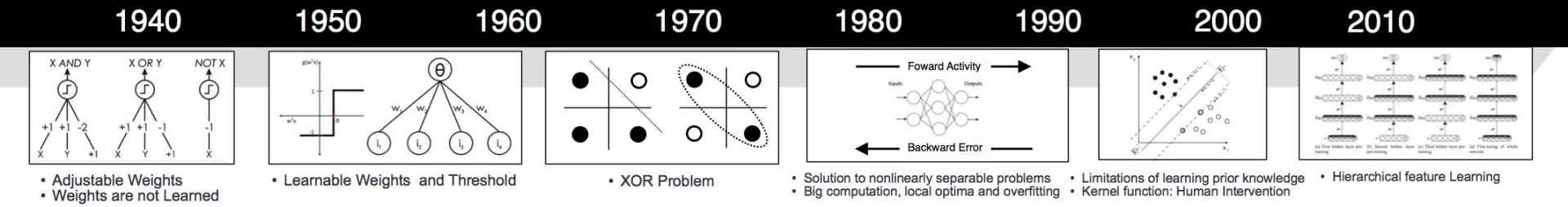


Deeper, Faster, and Smarter
—走向嵌入式深度学习时代

邓仰东
清华大学软件学院

提纲

1.背景

深度学习崛起和嵌入式应用需求

2.深度学习硬件

快速涌现的深度学习硬件

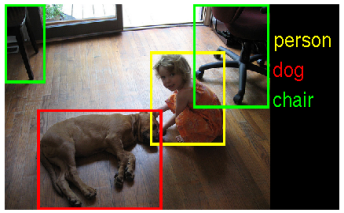
3.嵌入式深度学习

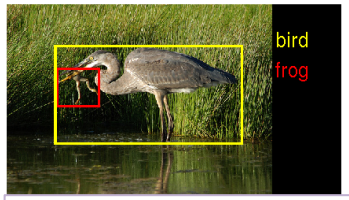
二值网络、深度压缩、加速器

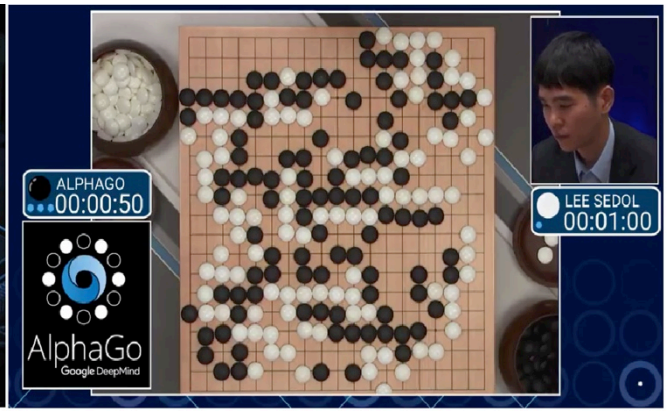
4.在研工作

走向能够思考和学习的机器

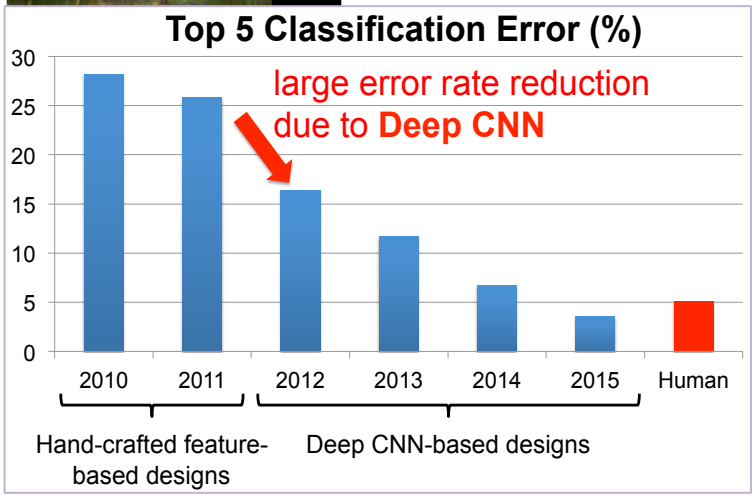
深度学习时代



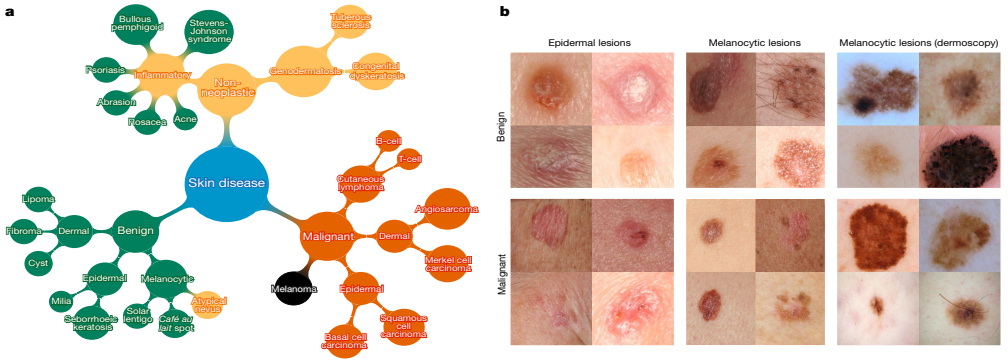




AlphaGo: Deep Reinforcement Learning for Go Game (Nature'16)



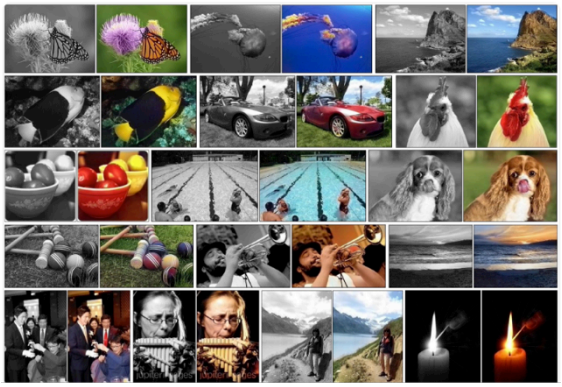
ImageNet: Image recognition, detection, and location



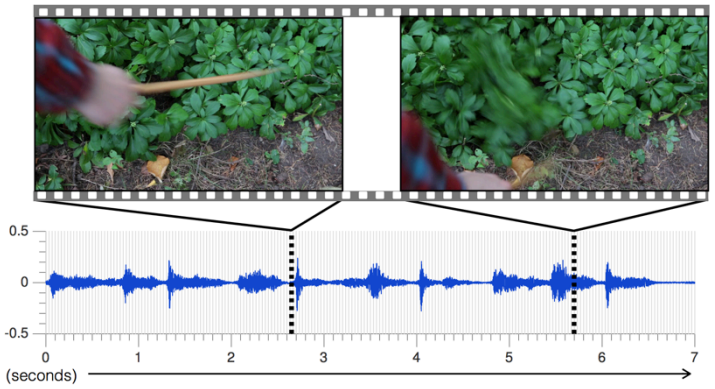
Dermatologist-level classification of skin cancer with deep neural networks (Nature'17)

深度学习应用

自动染色

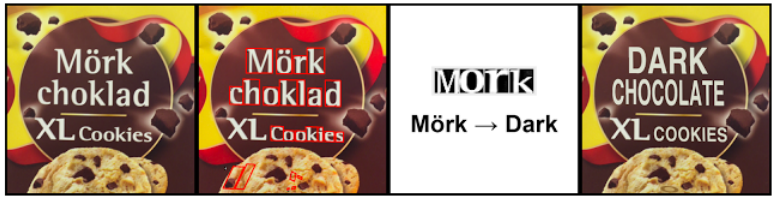


自动添加音响效果

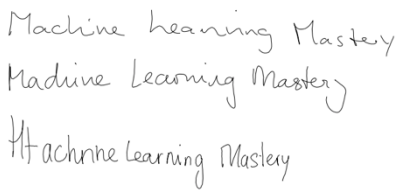


自动生成标题

翻译

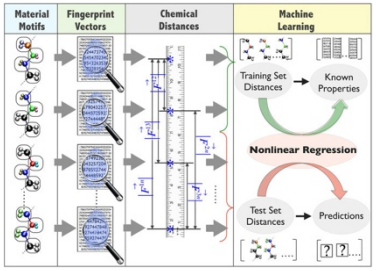


手写体生成

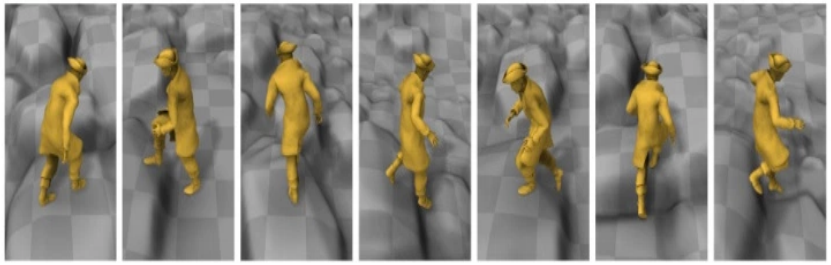


风格化

新材料发现



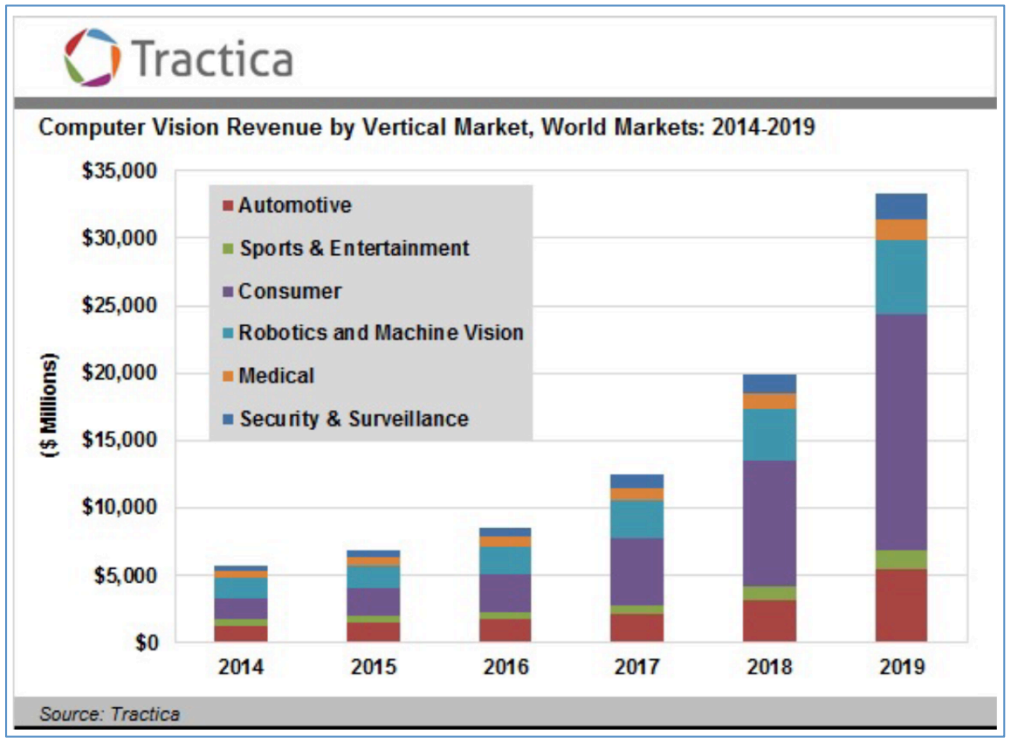
动画人物动作控制



计算机梦境

未来市场

- 深度学习技术在未来10年将形成5000亿美元的市场！



职业恐惧度

Probability of computerisation of different occupations, 2013
(1 = certain)

Job	Probability
Recreational therapists	0.003
Dentists	0.004
Athletic trainers	0.007
Clergy	0.008
Chemical engineers	0.02
Editors	0.06
Firefighters	0.17
Actors	0.37
Health technologists	0.40
Economists	0.43
Commercial pilots	0.55
Machinists	0.65
Word processors and typists	0.81
Real-estate sales agents	0.86
Technical writers	0.89
Retail salespeople	0.92
Accountants and auditors	0.94
Telemarketers	0.99

Source: "The Future of Employment: How Susceptible are Jobs to Computerisation?", by C. Frey and M. Osborne (2013)

Economist.com

局限

- 需要大量具有标签的样本数据
  - 但是实际应用中更多需要不间断学习
- 成功领域相对有限
  - ImageNet等数据集合的类别频率与真实世界不同
  - 但是较少解决实时、在线、控制类应用
- 与日常生活的融入性不足
  - 但是我们需要“不经意”的深度学习应用

FIRST YOU GET THE DATA, THEN YOU GET THE AI

APPLYING DEEP LEARNING TO REAL-WORLD PROBLEMS CAN BE MESSY

深度学习 + 嵌入式计算

嵌入式深度学习的挑战

- 存储器容量
  - ~100M权重参数
- 计算强度
  - 1-10G浮点乘加运算/inference
- 推断延迟
  - 10ms - 10s
- 能耗
  - ~10img/s/W

Metrics	LeNet-5	AlexNet	OverFeat (fast)	VGG-16	GoogLeNet (v1)	ResNet-50
Top-5 error	n/a	16.4	14.2	7.4	6.7	5.3
Input Size	28x28	227x227	231x231	224x224	224x224	224x224
# of CONV Layers	**2**	**5**	**5**	**16**	**21**	**49**
Filter Sizes	5	3, 5,11	3, 7	3	1, 3 , 5, 7	1, 3, 7
# of Channels	1, 6	3 - 256	3 - 1024	3 - 512	3 - 1024	3 - 2048
# of Filters	6, 16	96 - 384	96 - 1024	64 - 512	64 - 384	64 - 2048
Stride	1	1, 4	1, 4	1	1, 2	1, 2
# of Weights	26k	2.3M	16M	14.7M	6.0M	23.5M
# of MACs	1.9M	666M	2.67G	15.3G	1.43G	3.86G
# of FC layers	**2**	**3**	**3**	**3**	**1**	**1**
# of Weights	406k	58.6M	130M	124M	1M	2M
# of MACs	405k	58.6M	130M	124M	1M	2M
Total Weights	**431k**	**61M**	**146M**	**138M**	**7M**	**25.5M**
Total MACs	**2.3M**	**724M**	**2.8G**	**15.5G**	**1.43G**	**3.9G**

Network: GoogLeNet	Batch Size	Titan X (FP32)	Tegra X1 (FP32)	Tegra X1 (FP16)
Inference Performance	1	138 img/sec	33 img/sec	33 img/sec
Power		119.0 W	5.0 W	4.0 W
Performance/Watt		1.2 img/sec/W	6.5 img/sec/W	8.3 img/sec/W
Inference Performance	128 (Titan X) 64 (Tegra X1)	863 img/sec	52 img/sec	75 img/sec
Power		225.0 W	5.9 W	5.8 W
Performance/Watt		3.8 img/sec/W	8.8 img/sec/W	12.8 img/sec/W

提纲

1.背景

深度学习崛起和嵌入式应用需求

2.深度学习硬件

快速涌现的深度学习硬件

3.嵌入式深度学习

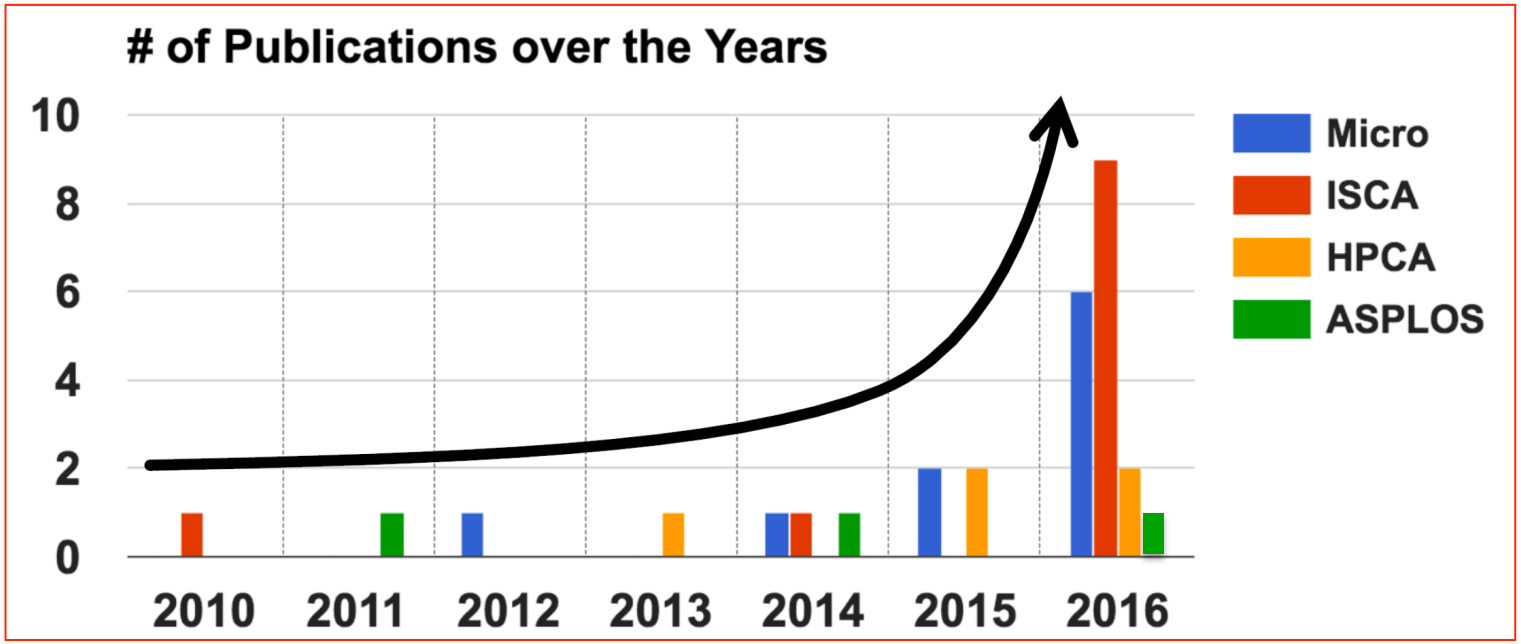
二值网络、深度压缩、加速器

4.在研工作

走向能够思考和学习的机器

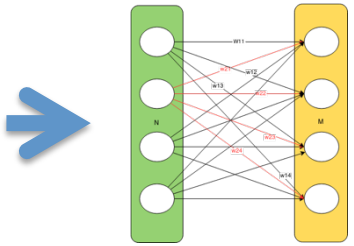
深度学习硬件研究快速兴起

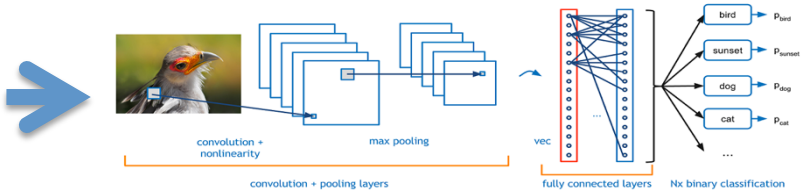
- ISCA, MICRO, HPCA, ASPLOS（计算机体系结构四大会议）
  - 2017年ISCA上6篇（6/54）
- 台积电年内将有>20颗深度神经网络芯片问世

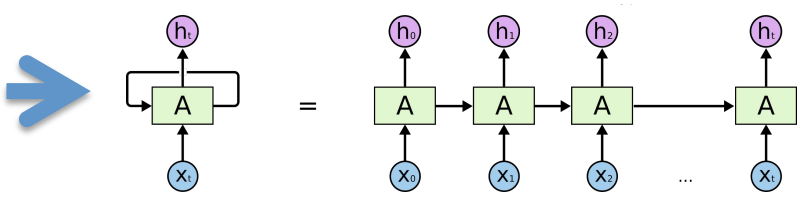


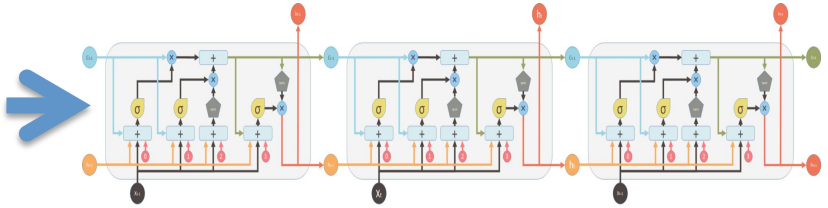
常见深度学习网络

- 全连接(Fully-Connected)网络
  - feed forward, a.k.a. multilayer perceptron (MLP)
- 卷积(Convolutional)神经网(CNN)
  - feed forward, sparsely-connected w/ weight sharing
- 递归(Recurrent)神经网(RNN)
  - feedback
- Long Short-Term Memory (LSTM)
  - feedback + Storage



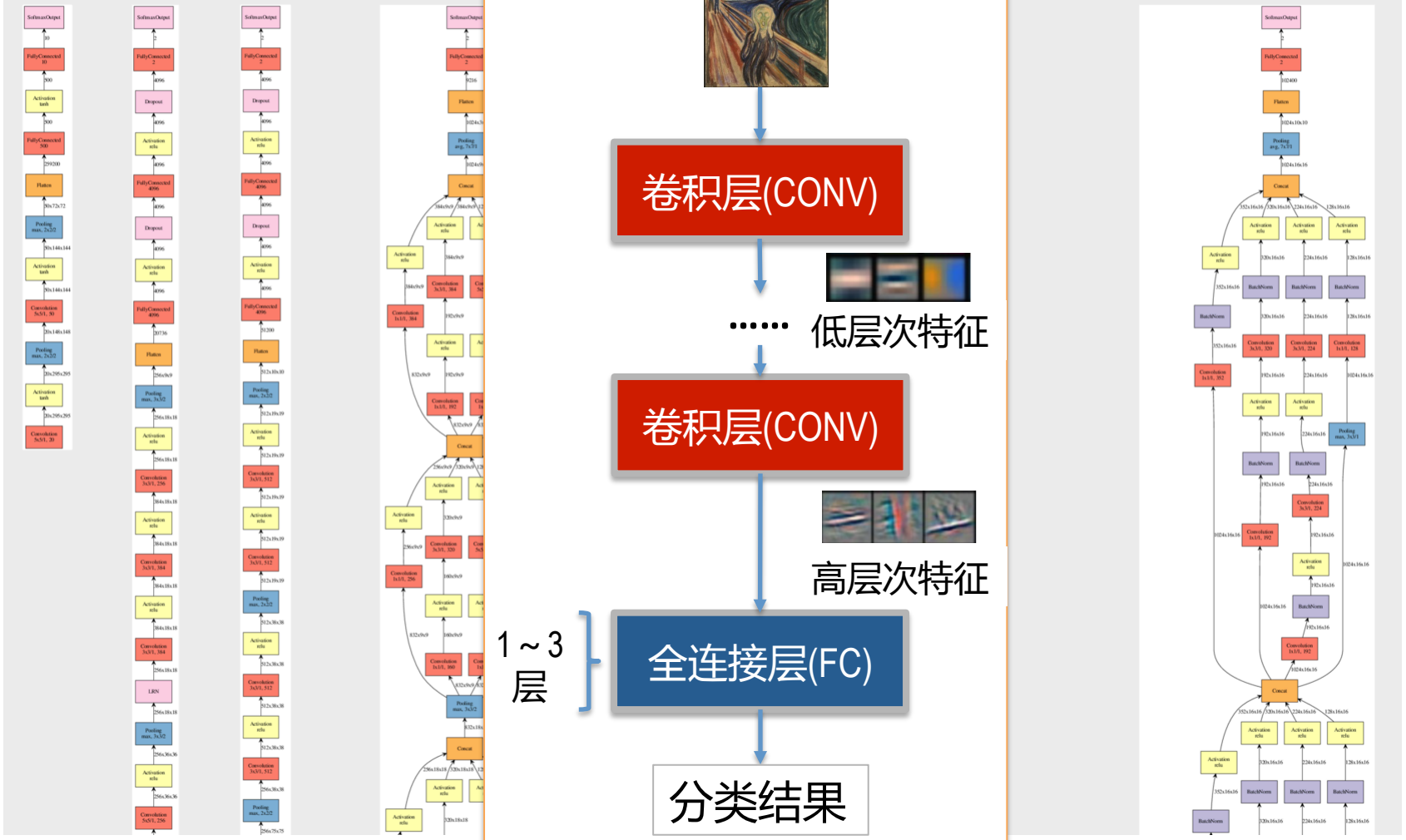






卷积神经网

LeNet AlexNet VGG GoogLe Inception BN



卷积层(CONV)

…… 低层次特征

卷积层(CONV)

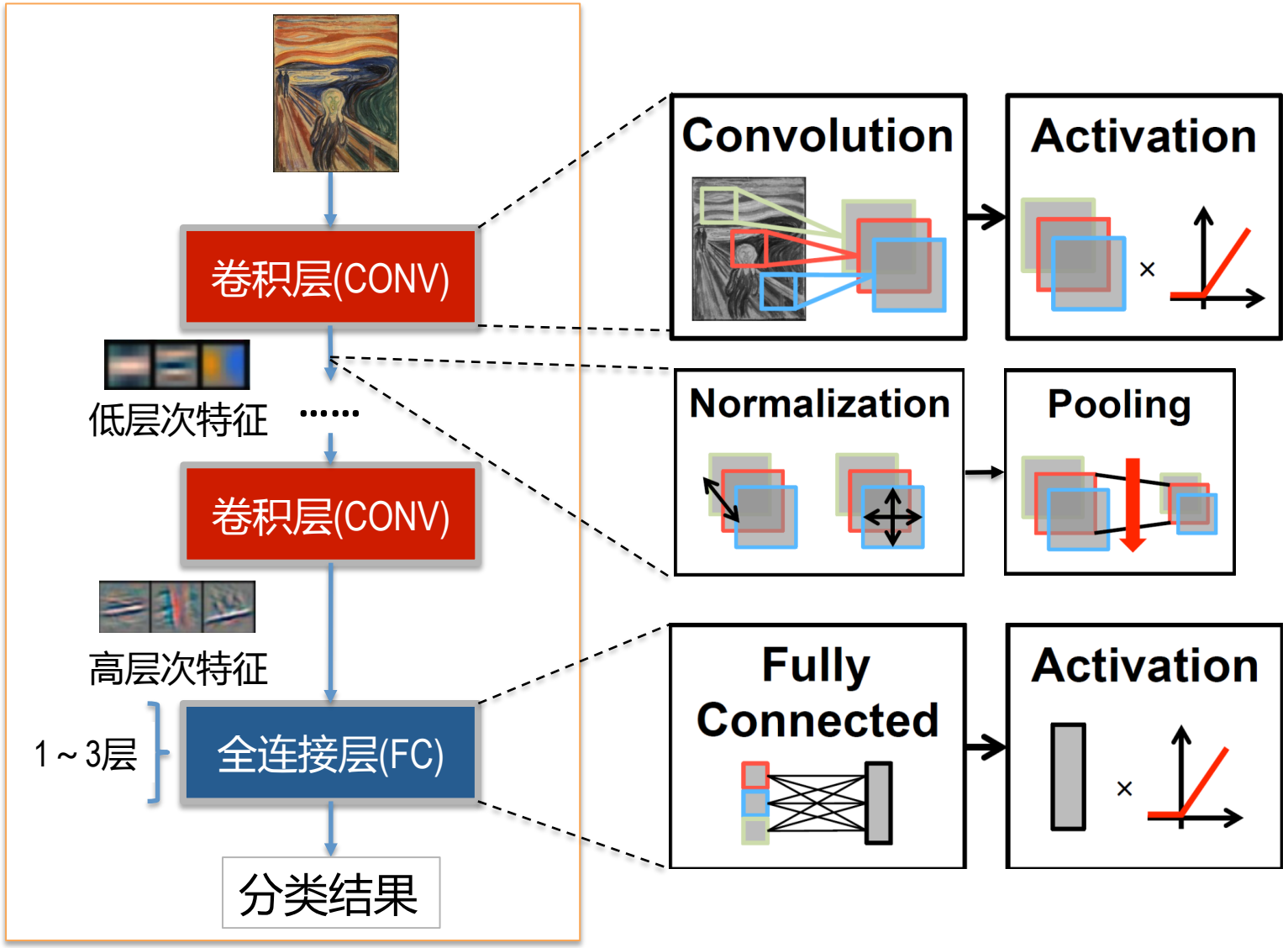
高层次特征

1 ~ 3 层

全连接层(FC)

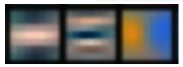
分类结果

卷积神经网


卷积层(CONV)
低层次特征
……
卷积层(CONV)
高层次特征
1 ~ 3层
全连接层(FC)
分类结果
Convolution
Activation
Normalization
Pooling
Fully Connected
Activation

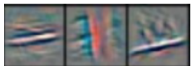
卷积神经网

卷积层(CONV)



低层次特征 ……

卷积层(CONV)



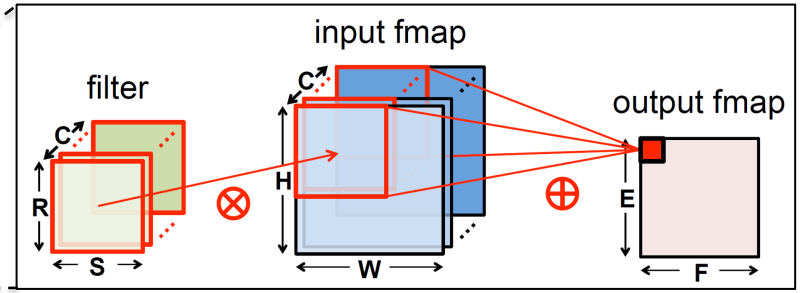
高层次特征

1 ~ 3层

全连接层(FC)

分类结果

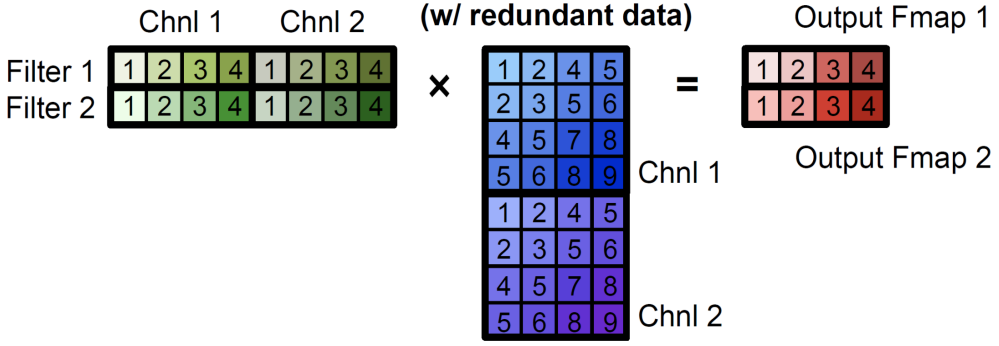
卷积运算占据了90%的计算量和能耗！



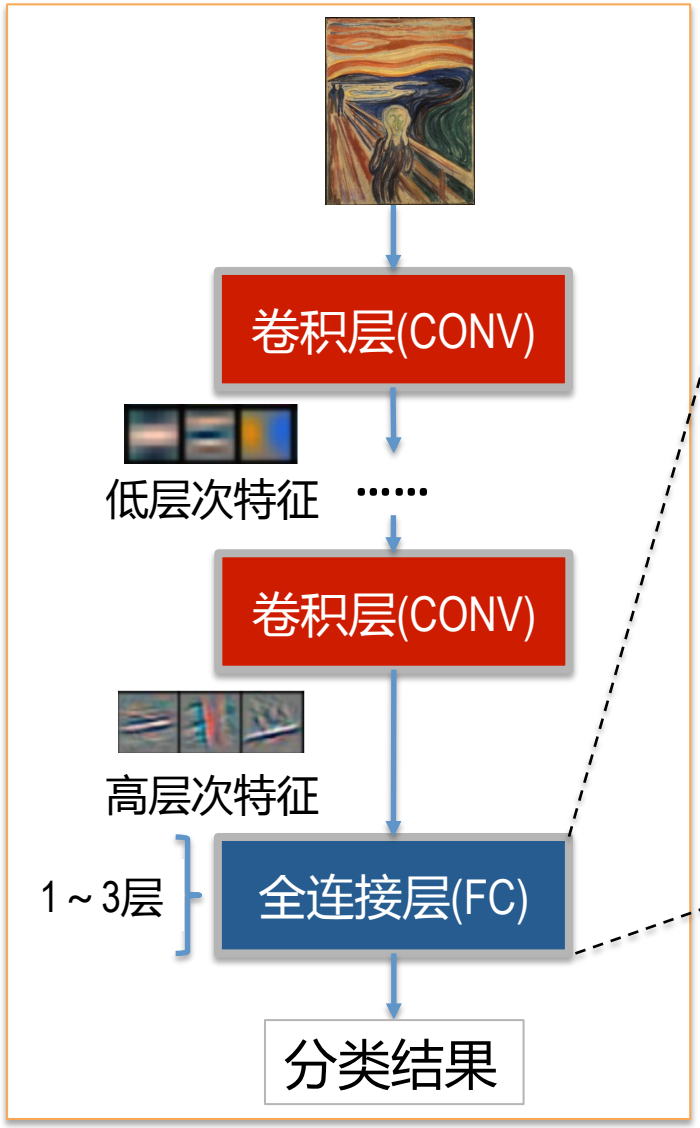
Output fmaps **Biases** **Input fmaps** **Filter weights**

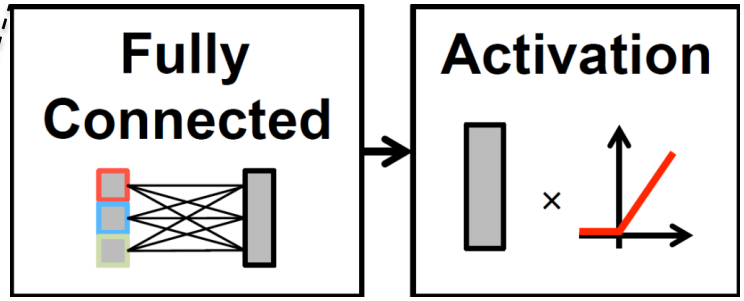
$$\mathbf{O}[n][m][x][y] = \text{Activation}(\mathbf{B}[m] + \sum_{i=0}^{R-1}\sum_{j=0}^{S-1}\sum_{k=0}^{C-1} \mathbf{I}[n][k][Ux+i][Uy+j] \times \mathbf{W}[m][k][i][j]),$$

Toeplitz Matrix (w/ redundant data)

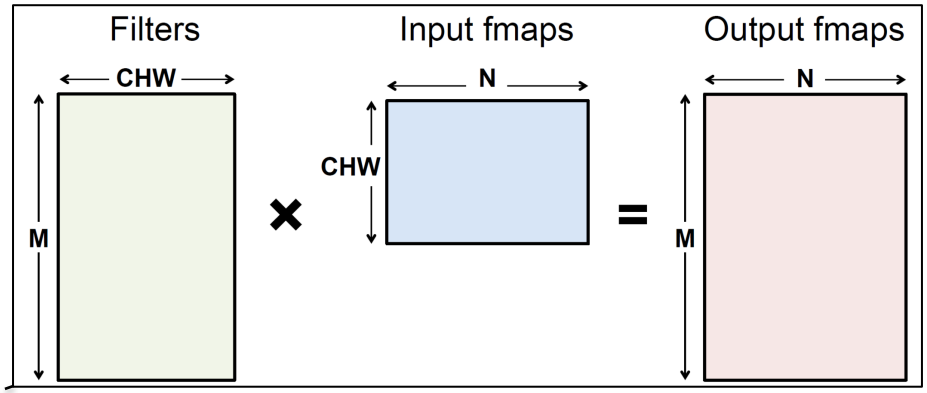


全连接层

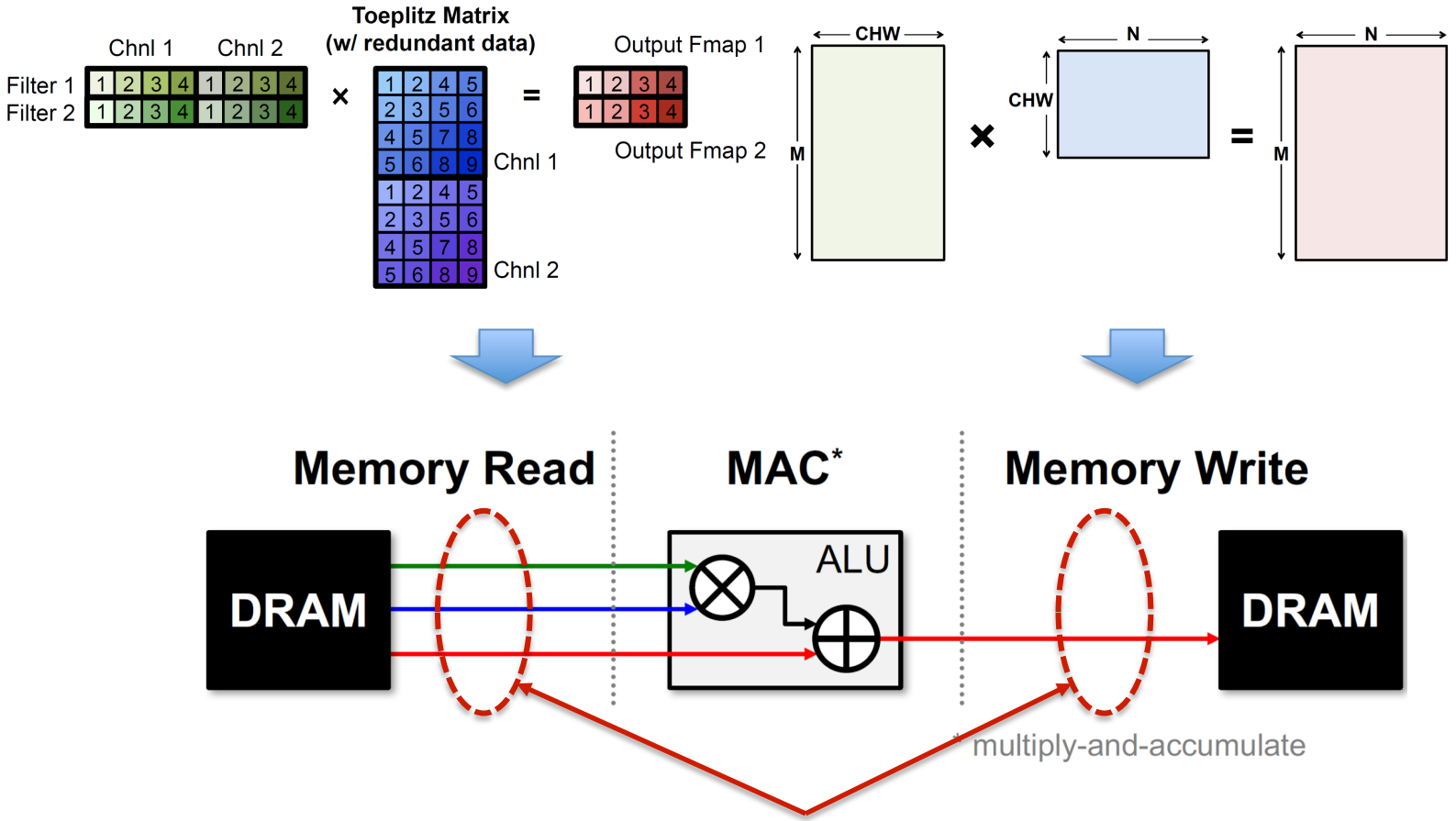




全连接层的乘法运算可以占据50%以上FPGA硬件资源！



卷积神经网计算数据流

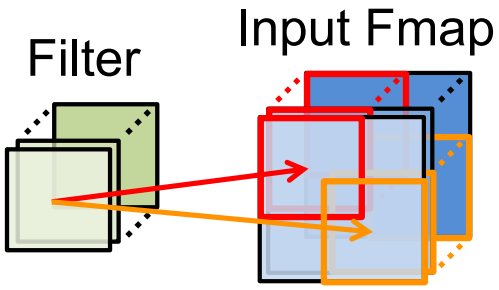


瓶颈！（e.g., AlexNet需要30亿次DRAM访问）

数据重用

卷积核参数的重用

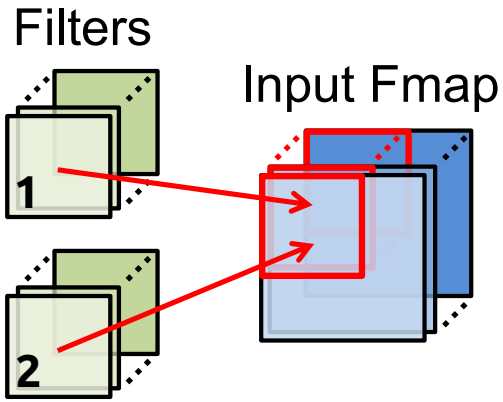
CONV layers only
(sliding window)



Reuse: Activations
Filter weights

Feature Map的重用

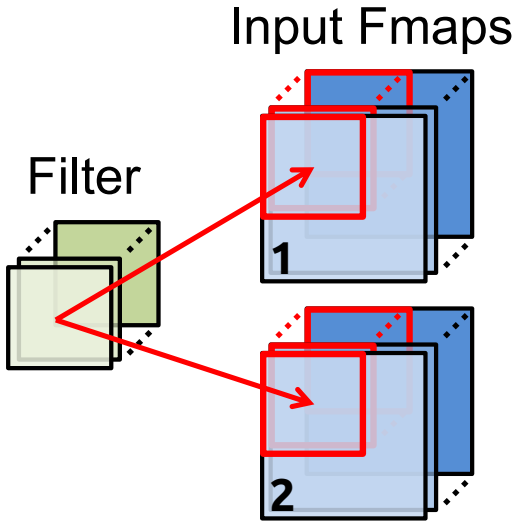
CONV and FC layers



Reuse: Activations

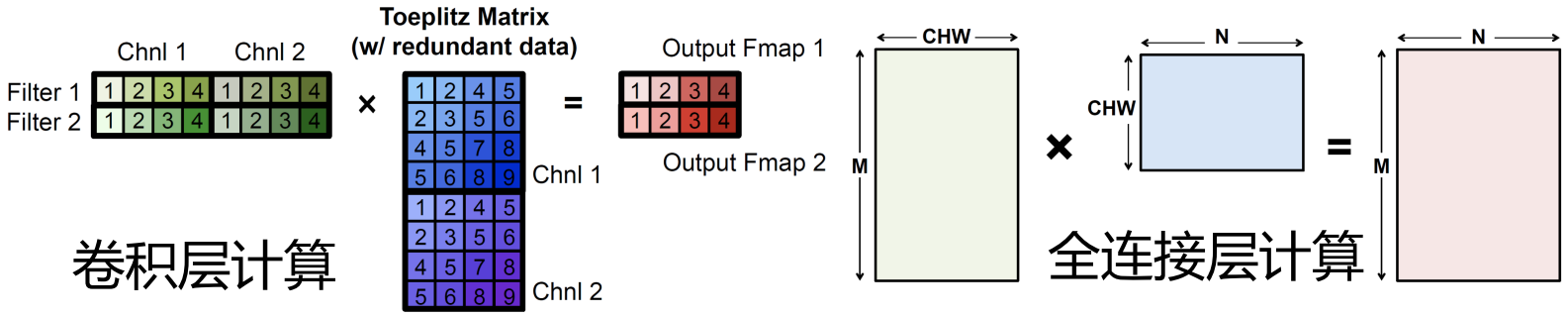
卷积核参数在多张图片的重用

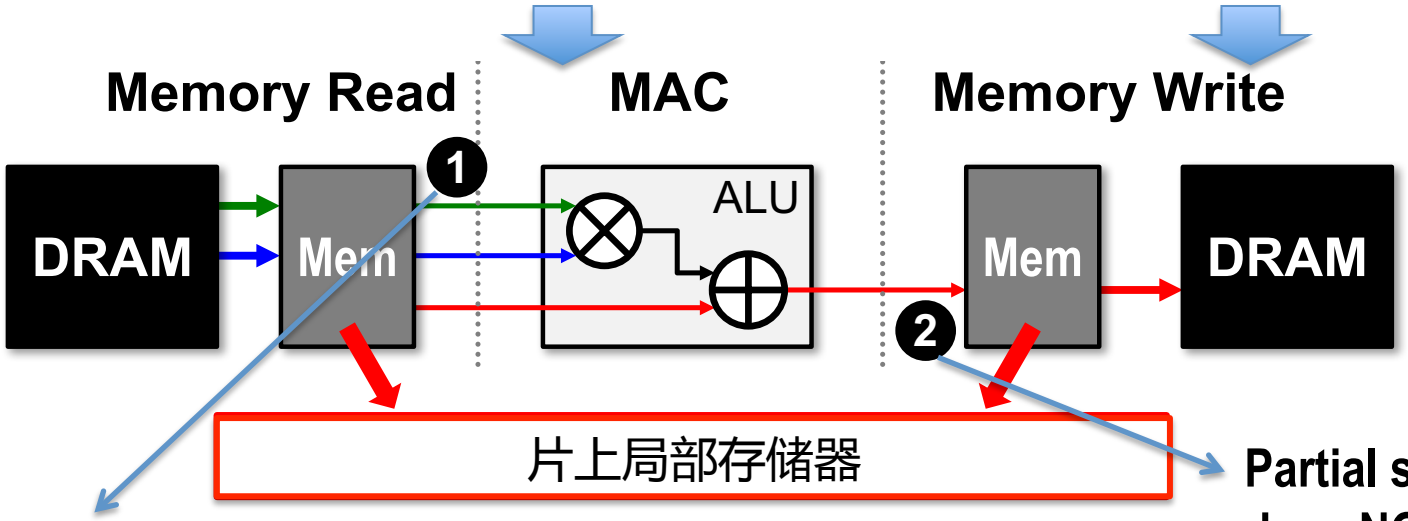
CONV and FC layers
(batch size > 1)



Reuse: Filter weights

局部存储器

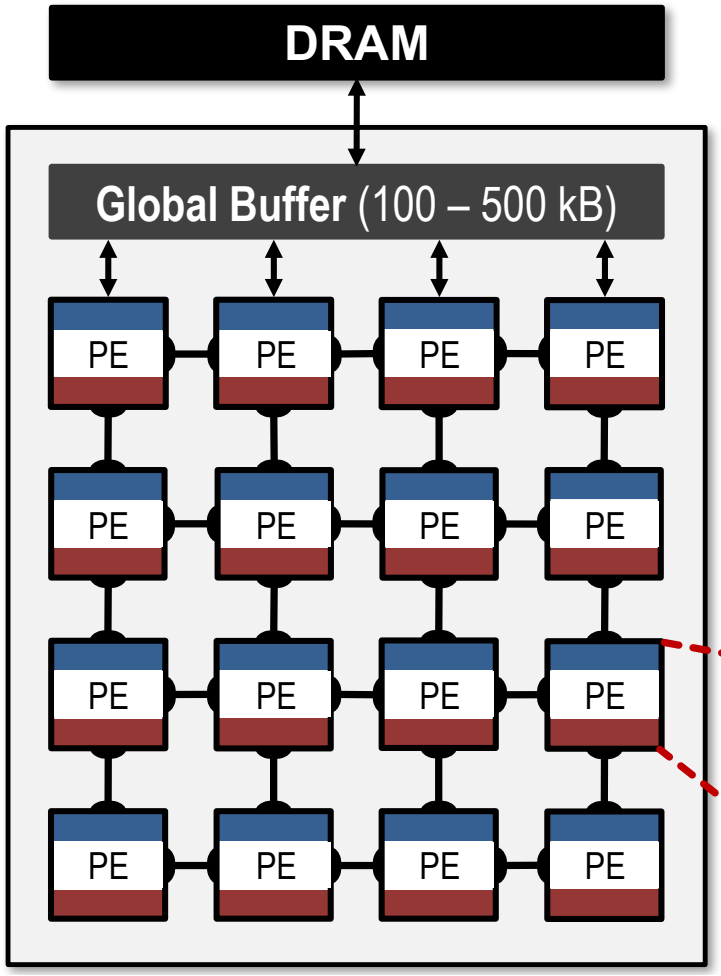




Can reduce DRAM reads of filter/fmap by up to 500×

Partial sum accumulation does NOT have to access DRAM

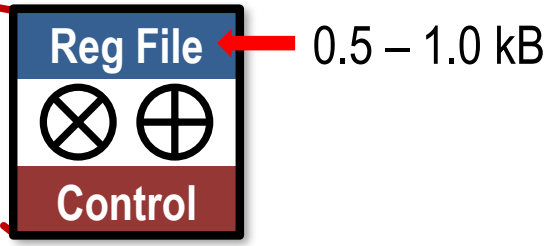
典型卷积神经网硬件体系结构



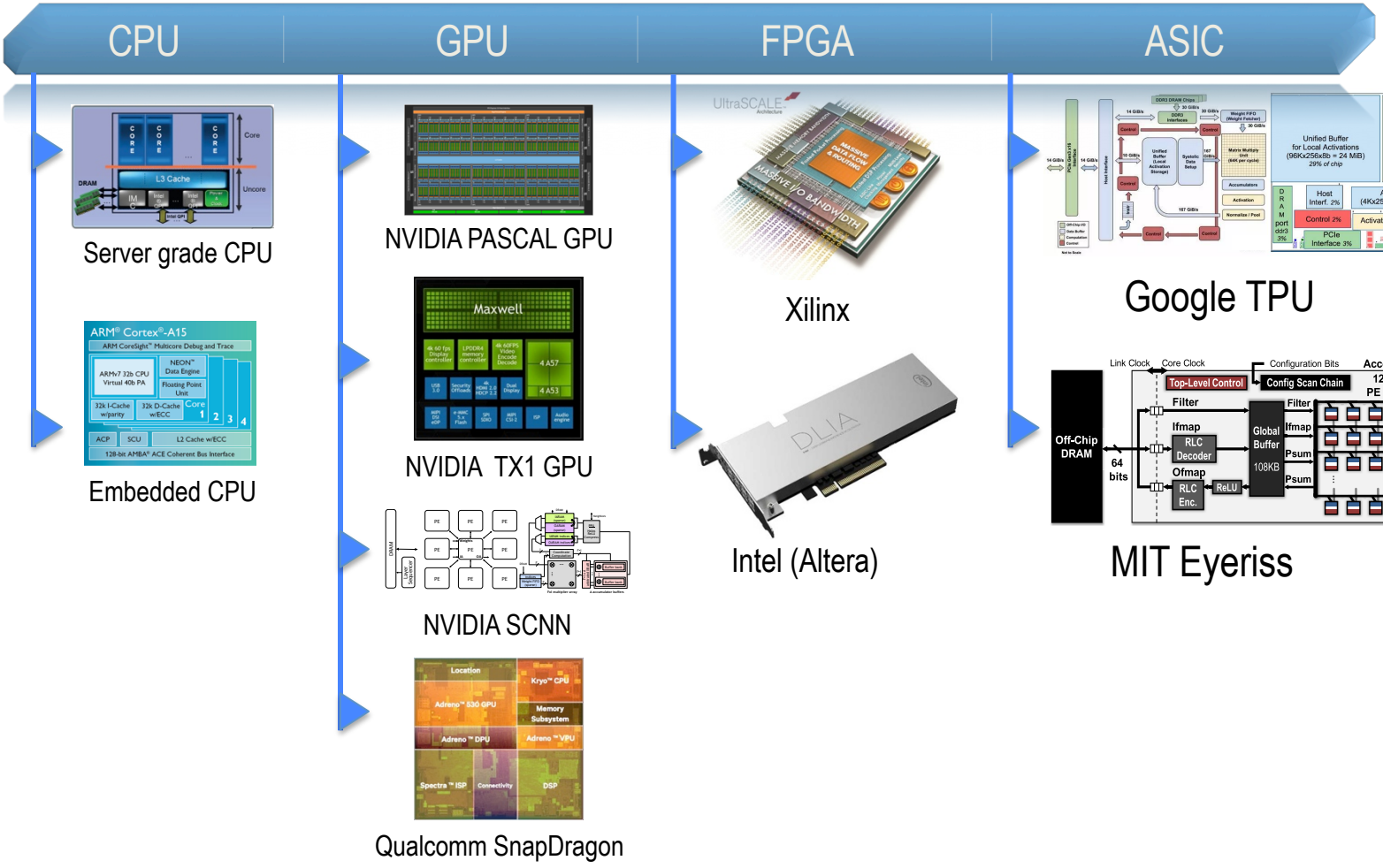
Local Memory Hierarchy

- Global Buffer
- Direct inter-PE network
- PE-local memory (RF)

Processing Element (PE)



卷积神经网硬件


CPU
GPU
FPGA
ASIC
Server grade CPU
Embedded CPU
NVIDIA PASCAL GPU
NVIDIA TX1 GPU
NVIDIA SCNN
Qualcomm SnapDragon
Xilinx
Intel (Altera)
Google TPU
MIT Eyeriss

提纲

1.背景

深度学习崛起和嵌入式应用需求

2.深度学习硬件

快速涌现的深度学习硬件

3.嵌入式深度学习

二值网络、深度压缩、加速器

4.在研工作

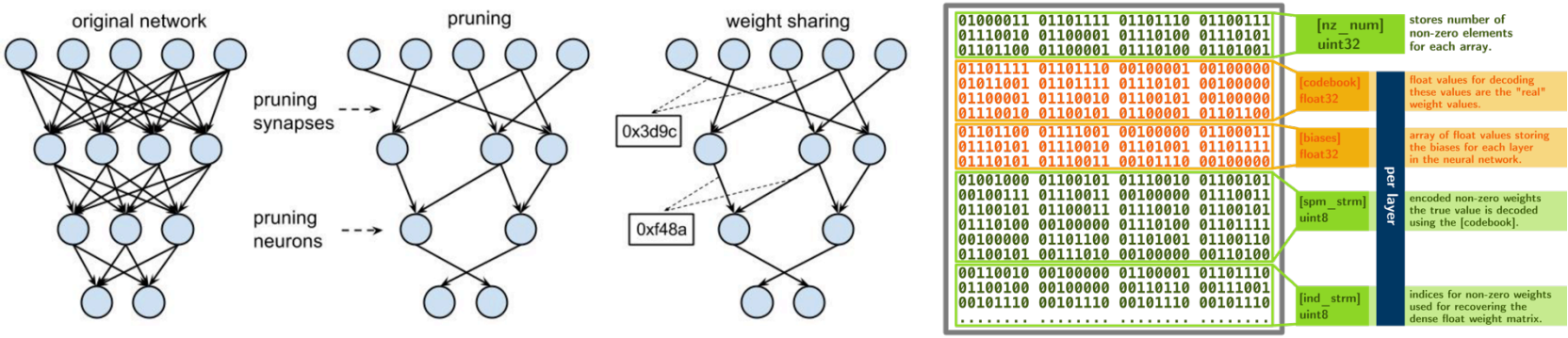
走向能够思考和学习的机器

二值深度神经网络

Data set	MNIST	SVHN	CIFAR-10
Binarized activations+weights, during training and test			
BNN (Torch7)	1.40%	2.53%	10.15%
BNN (Theano)	0.96%	2.80%	11.40%
Committee Machines' Array (Baldassi et al., 2015)	1.35%	-	-
Binarized weights, during training and test			
BinaryConnect (Courbariaux et al., 2015)	1.29± 0.08%	2.30%	9.90%
Binarized activations+weights, during test			
EBP (Cheng et al., 2015)	2.2± 0.1%	-	-
Bitwise DNNs (Kim & Smaragdis, 2016)	1.33%	-	-
Ternary weights, binary activations, during test			
(Hwang & Sung, 2014)	1.45%	-	-
No binarization (standard results)			
Maxout Networks (Goodfellow et al.)	0.94%	2.47%	11.68%
Network in Network (Lin et al.)	-	2.35%	10.41%
Gated pooling (Lee et al., 2015)	-	1.69%	7.62%

*Courbariaux et al., Binarized Neural Networks: Training Neural Networks withWeights and Activations Constrained to +1 or -1, 2016

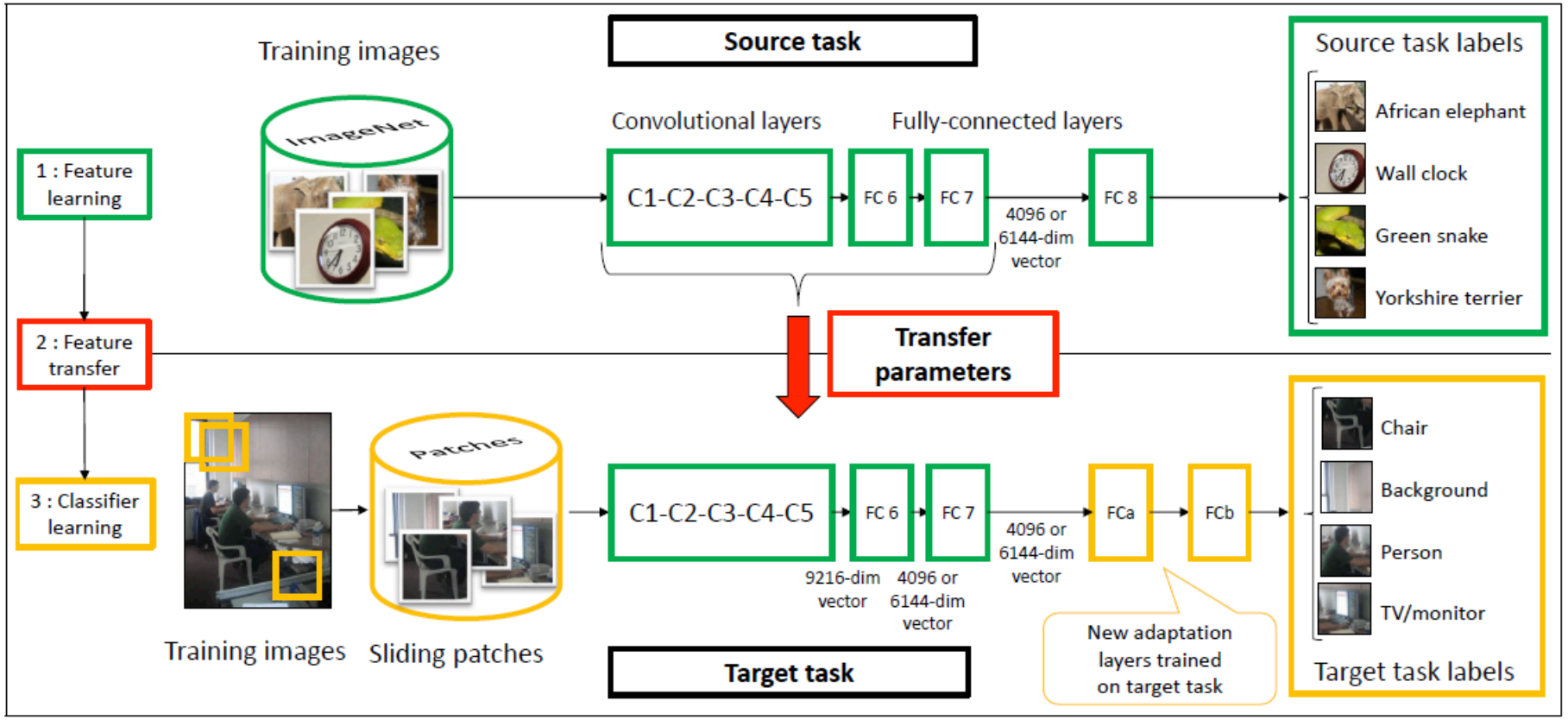
深度压缩



Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
Lenet-300-100	1070KB	27KB	**40x**	98.36%	98.42%
Lenet-5	1720KB	44Kb	**39x**	99.20%	99.26%
AlexNet	240MB	6.9MB	**35x**	80.27%	80.30%
VGGNet	550MB	11.3MB	**49x**	88.68%	89.09%
GoogleNet	28MB	2.8MB	**10x**	88.90%	88.92%
SqueezeNet	4.8MB	0.47MB	**10x**	80.32%	80.35%

*Han et al., Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, 2016

转移学习


Training images
Source task
ImageNet
Convolutional layers
Fully-connected layers
1 : Feature learning
C1-C2-C3-C4-C5
FC 6
FC 7
4096 or 6144-dim vector
FC 8
Source task labels
African elephant
Wall clock
Green snake
Yorkshire terrier
2 : Feature transfer
Transfer parameters
3 : Classifier learning
Patches
C1-C2-C3-C4-C5
FC 6
FC 7
4096 or 6144-dim vector
FCa
FCb
9216-dim vector
4096 or 6144-dim vector
Chair
Background
Person
TV/monitor
Training images
Sliding patches
Target task
New adaptation layers trained on target task
Target task labels

嵌入式深度学习

- 大量深度神经网络芯片即将出现
  - 大多针对嵌入式设备
- 目前典型应用只在嵌入式设备上进行推断
  - 训练离线进行（但已经出现基于FPGA的训练加速器）
    - 使用转移学习技术适应新的应用领域
  - 使用二值网络和深度压缩降低存储容量和计算速度
- 未来重点在于融合
  - 多模态数据融合实现复杂环境认知
  - 与贝叶斯推理技术融合形成高层次认知能力
  - 与在线学习和增强式学习技术融合形成不间断学习能力

提纲

1.背景

深度学习崛起和嵌入式应用需求

2.深度学习硬件

快速涌现的深度学习硬件

3.嵌入式深度学习

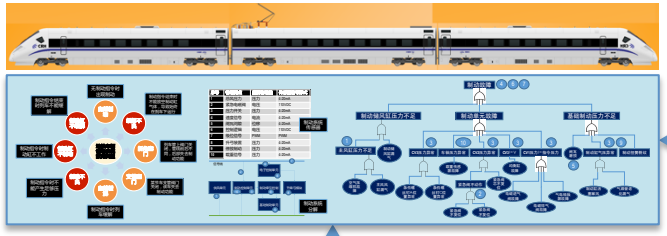
二值网络、深度压缩、加速器

4.在研工作

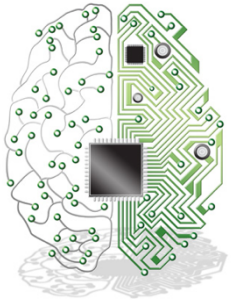
走向能够思考和学习的机器

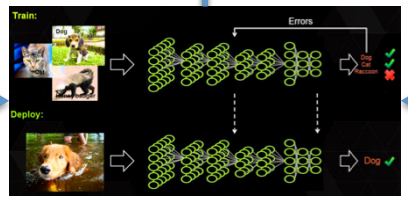
在研工作

工业装备故障预测和健康管理
(自然科学基金重大仪器项目、中车集团重大专项)

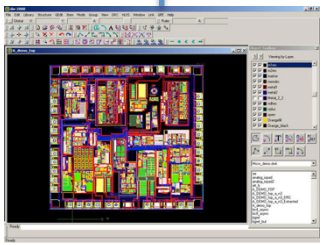


基于随机采样计算的类脑计算机
(Xilinx合作项目)

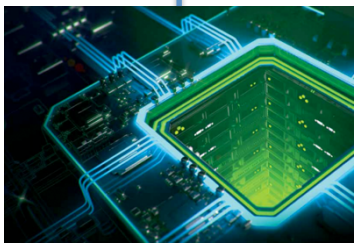




深度学习
Pascal VOC Challenge物体检测
当前排名第1



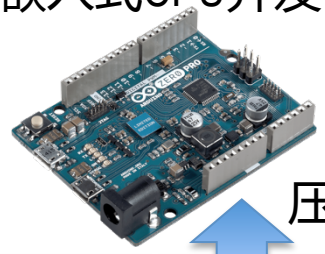
集成电路计算机辅助设计
(Intel, 清华大学骨干人才支持计划)



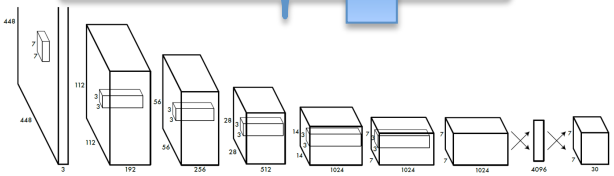
GPU计算
自然科学基金，NVIDIA合作教授奖,
ICCD'13最佳论文奖

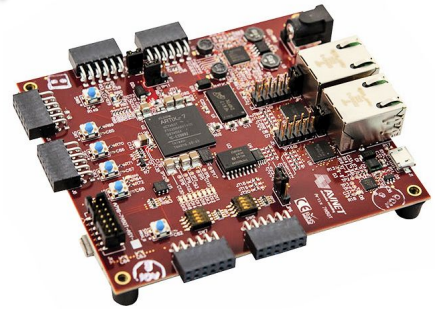
在研项目1：专业人员行为识别

嵌入式CPU开发板



压缩





基于FPGA的嵌入式设备

在研项目2：轨道装备预测式健康管理

全面提高运行安全和效率

全面优化修程和修制

全面提升维修水平和出勤率

重大故障在线预警

部件剩余寿命估计

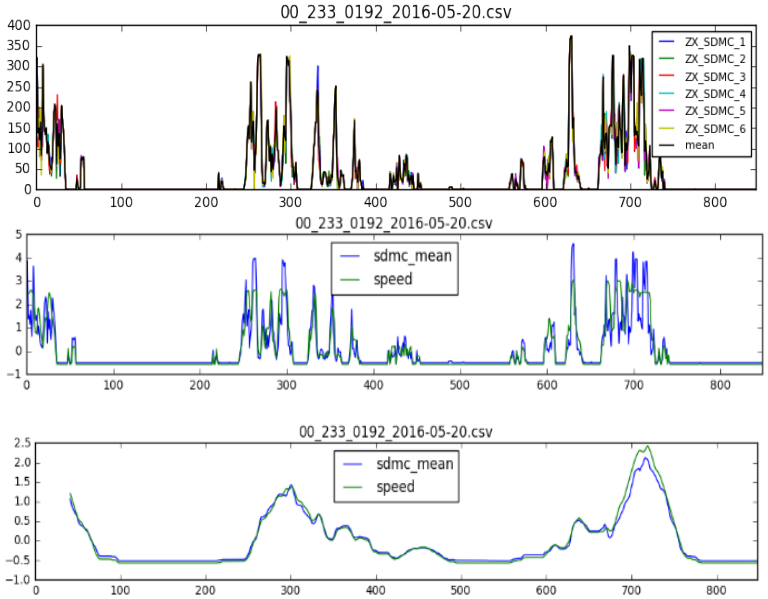
自动化故障诊断和维修

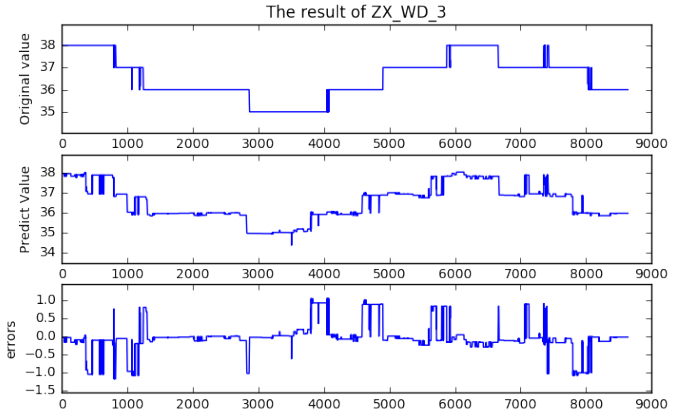
利用中国铁路特有的、全球全球最丰富环境的运营样本、数据和经验，大幅度提高运维效率、降低维修成本，支撑铁路可持续发展

合作单位：中车集团、丰台机务段、铁总信息中心

成果：列车走行部轴温分析和预测

基于深度神经网络的机理挖掘和轴温预测算法





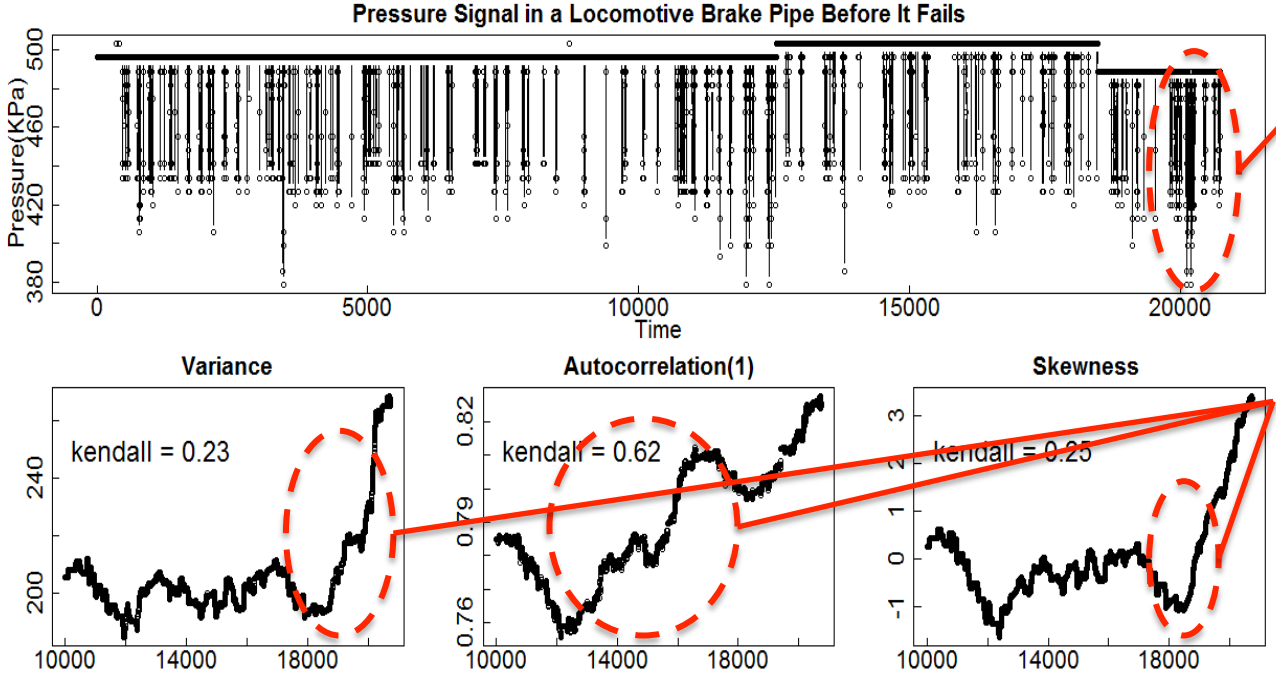
$$T = 8.45 + 0.407*ZX\_WD\_1 + 0.152*ZX\_HW\_1 + 0.110*ZX\_HW\_2 + 0.0309*ZX\_WD\_3 + 0.000971*ZD\_SPEED + 0.000796*ZD\_LLJ$$

- 目的
  - 推断物理机理和确定主导影响因素（故障分析）
  - 预测轴承温度变化趋势

*数据来源：CMD数据

成果：机车刹车管故障预测

基于深度神经网络的时间序列早期分类算法



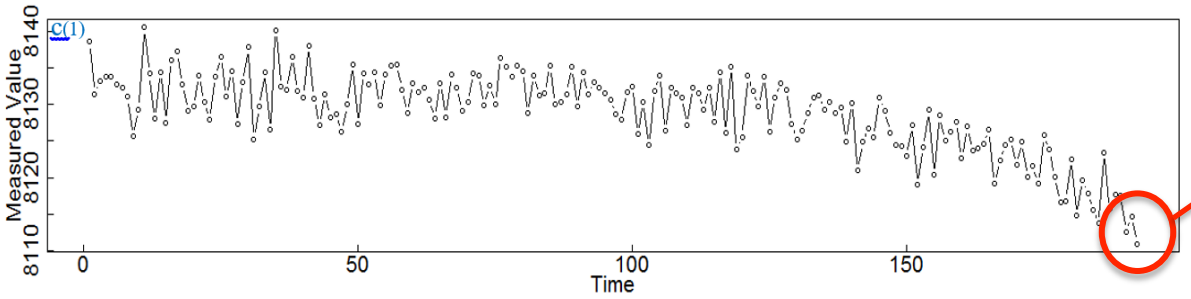
20500秒时刹车管失效

使用关键跳变检测算法在18000秒左右发现异常数据模式

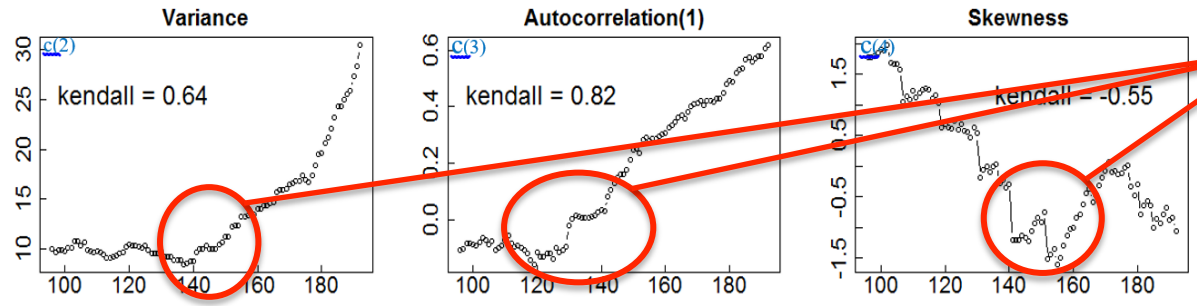
- 数据来源：苏家屯机务段某HXN3机车
- 数据描述：反映从正常运行到故障发生过程的风机气压、管压等状态数据

成果：涡轮风扇发动机故障预测

基于深度神经网络的时间序列早期分类算法



196分钟发生故障导致发动机失速（停转）



使用关键跳变检测算法在150分钟左右发现异常数据模式

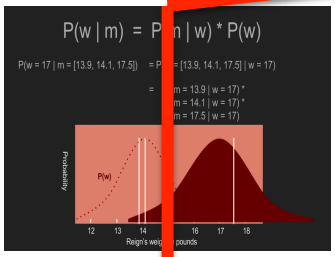
- 数据来源：NASA涡轮风扇航空发动机数据
- 数据描述：反映从正常运行到故障发生过程的气压、温度等状态数据

在研项目3：类脑贝叶斯计算机

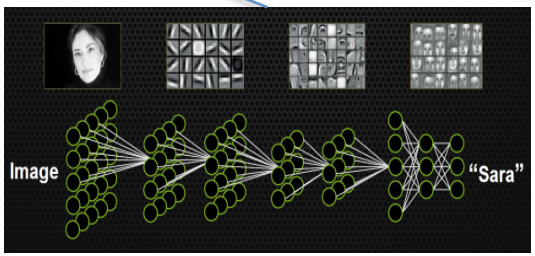
Computational Level

Human cognition problems

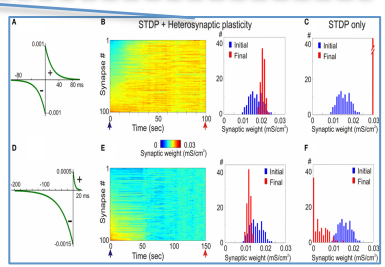
Algorithmic Level



Bayesian

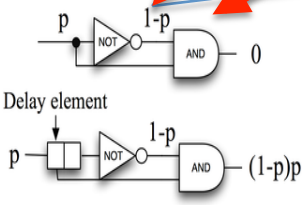


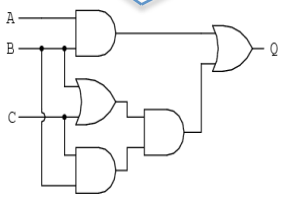
Deep Neural Network



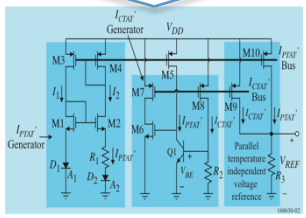
STDP Learning

Implementational Level

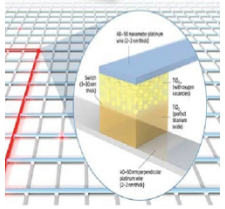




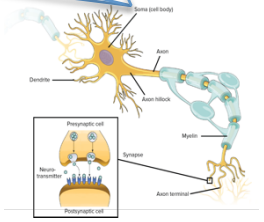
Digital circuits



Analog circuits



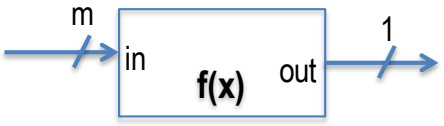
Memristor



Neuron

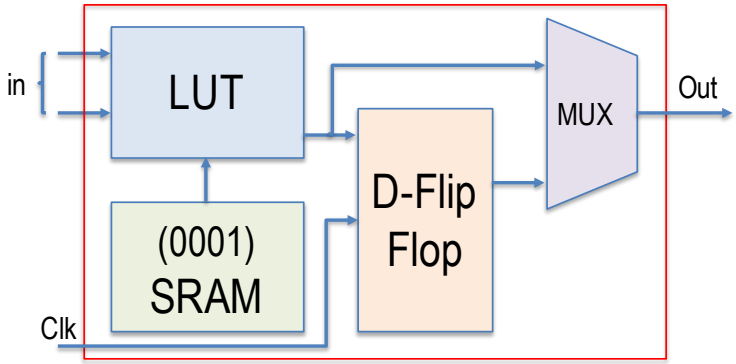
在研项目3：类脑贝叶斯计算机

- A Boolean logic gate (LUT here) is characterized by its truth table

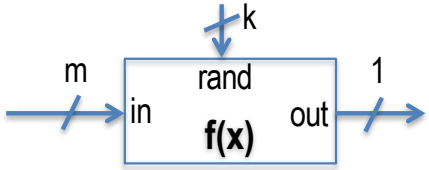


in	out
00	0
01	0
10	0
11	1

f(in) = AND(in1, in2)



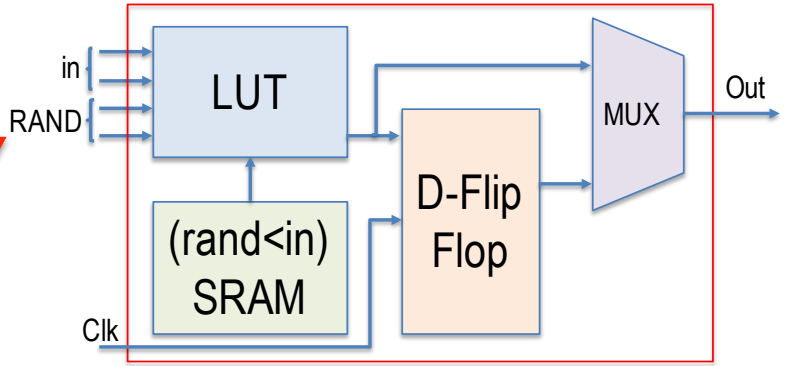
- A stochastic gate is characterized by its conditional probability table



RAND	out	P
0000	0	1
	1	0
…	…	…
0111	0	1/2
	1	1/2

f(in) = rand < in i.e., P(out | in)

Source of entropy



在研项目3：类脑贝叶斯计算机